Frequentist Methods in Inverse Problems

Sandia CSRI Workshop on Large-Scale Inverse Problems and Quantification of Uncertainty
10-12 September 2007

P.B. Stark

Department of Statistics

University of California

Berkeley, CA 94720-3860

statistics.berkeley.edu/~stark

Abstract

"Inverse problems" lives at the confluence of applied math, substantive fields in science and engineering, and statistics.

To quantify uncertainty statistically, probability has to enter the problem from somewhere. Where?

Ideas from statistical decision theory provide a useful perspective. Some statistical notions are closely related to deterministic ones.

Notions of optimality differ; measures of performance interesting. Helps to look at methods from many angles.

Prior information crucial. Prior probability distributions usually add more info than constraints—harder to justify.

Bayesian framework, while appealing and flexible, often leaves the realm of "Science."

Nature this week:

[an apparent surge in the impact flux of km-sized bodies] was triggered by the catastrophic disruption of the parent body of the asteroid Baptistina, which we infer was a ~170-km-diameter body [] that broke up 160 Myr ago []. We find that this asteroid shower is the most likely source (>90 per cent probability) of the Chicxulub impactor that produced the Cretaceous/Tertiary (K/T) mass extinction event 65 Myr ago.

Bottke, W.F., D. Vokrouhlický & D. Nesvorný, 2007. An asteroid breakup 160 Myr ago as the probable source of the K/T impactor, *Nature*, **449**, 48–53.

What (in heavens) does this mean?

This is an inverse problem—what is the source of the K/T impactor?

Solution is stated as a probability.

How is “probability” to be interpreted?

- What ties the number to the world we live in?
- How is the number calculated?
- Where does “randomness” enter the problem?

Several “theories of probability.”

All have the same axioms; interpretations differ.

Coin Tosses. What does P (heads) = 1/2 mean?

- Equally likely outcomes: Nature indifferent; principle of insufficient reason
- Frequency theory: long-term limiting relative frequency
- Subjective theory: strength of belief
- Probability models: property of math model; testable predictions
- Math coins - real coins.
- Weather predictions: look at sets of assignments. Scoring rules.

Littlewood, 1953

Mathematics (by which I shall mean pure mathematics) has no grip on the real world; if probability is to deal with the real world it must contain elements outside mathematics; the meaning of ‘probability’ must relate to the real world, and there must be one or more ‘primitive’ propositions about the real world, from which we can then proceed deductively (i.e. mathematically). We will suppose (as we may by lumping several primitive propositions together) that there is just one primitive proposition, the ‘probability axiom,’ and we will call it A for short. Although it has got to be true, A is by the nature of the case incapable of deductive proof, for the sufficient reason that it is about the real world . . . .

Littlewood, contd.

There are 2 schools. One, which I will call mathematical, stays inside mathematics, with results that I shall consider later. We will begin with the other school, which I will call philosophical. This attacks directly the ‘real’ probability problem; what are the axiom A and the meaning of ‘probability’ to be, and how can we justify A? It will be instructive to consider the attempt called the ‘frequency theory’. It is natural to believe that if (with the natural reservations) an act like throwing a die is repeated n times the proportion of 6’s will, with certainty, tend to a limit, p say, as $n \to \infty$. (Attempts are made to sublimate the limit into some Pickwickian sense—‘limit’ in inverted commas.

Littlewood, contd.

But either you mean the ordinary limit, or else you have the problem of explaining how ‘limit’ behaves, and you are no further. You do not make an illegitimate conception legitimate by putting it into inverted commas.) If we take this proposition as ‘A’ we can at least settle off-hand the other problem, of the meaning of probability; we define its measure for the event in question to be the number p. But for the rest this A takes us nowhere. Suppose we throw 1000 times and wish to know what to expect. Is 1000 large enough for the convergence to have got under way, and how far? A does not say. We have, then, to add to it something about the rate of convergence. Now an A cannot assert a certainty about a particular number n of throws, such as

Littlewood, contd.

‘the proportion of 6’s will certainly be within p±ε for large enough n (the largeness depending on ε)’. It can only say ‘the proportion will lie between p±ε with at least such and such probability (depending on ε and n_0) whenever $n > n_0$ ’. The vicious circle is apparent. We have not merely failed to justify a workable A; we have failed even to state one which would work if its truth were granted. It is generally agreed that the frequency theory won’t work. But whatever the theory it is clear that the vicious circle is very deep-seated: certainty being impossible, whatever A is made to state can be stated only in terms of ‘probability’.

Bottke et al., 2007 claim

P(Chicxulub impactor was from Baptistina breakup) ³ 0.9

- What does that mean?
  - None of the standard theories helps.
- Where does the number come from?
  - It comes from a complex model.
  - Many assumptions and data processing steps.

Some of the steps

- Start with catalog of estimated orbital parameters
- Apply a hierarchical clustering algorithm to the catalog, using several ad hoc tuning parameters and an ad hoc choice of a metric. After some fiddling and judgment calls, classify each asteroid as belonging to BAF or not.
- Use Sloan Digital Sky Survey color data to refine the classification; ad hoc choice of a threshold velocity. Lingering questions about left censoring, etc.
- Estimate age of BAF using parametric model for how the center of an asteroid family gets depleted as the family evolves; perform repeated Monte Carlo using Gaussian distributions for the components of the parent body's velocity, with same SD in all directions. Part of the relationship between SD and diameter calibrated to data from a 5.8Myr asteroid family. Magnitude converted to size using rule of thumb assuming constant geometric albedo. Choose orientation of spin axes s.t. $\cos \varepsilon \sim U[-1,1]$ (error in ms.); rotational velocity truncated Gaussian.

More steps

- Track fragments, use simple model for thermal effects on the orbit and assumed constant values for thermal conductivity, specific heat capacity and density.

- Evolve obliquity and rotation rate using simple DEQ model for YORP effect—with a fudge factor.

- Model how collisions affect spin vector of asteroids: some collisions give a new random spin.

- Fit free parameters within a priori intervals; reject initial conditions if “best fit” is poor. Result:
  age T = 160MYr (+30/-20)
  scalar velocity V_0 = 40±10 m/s

- Start with a compilation of estimates of magnitudes of objects classified as BAF. Adjust magnitudes to undo bias from an assumption the surveys used to calculate magnitude. Take magnitudes for objects with larger than threshold orbital semimajor axis; assume these are a mirror image, so double the counts to get a magnitude distribution for BAF.

and more...

- Make a parametric bootstrap of magnitude values with Gaussian errors. Truncate the data at magnitude 15.3. Fit power-law model to 10,000 pseudo datasets. Extrapolate the model to magnitude 19.1, corresponding to 1km diameter objects. Take mean and SD: $(1.36 \pm 0.3) \times 10^5$. Get size frequency distribution this way.

- Estimate loss of objects from BAF by tracking test members using numerical dynamical model for 70 5-km asteroids, assumed density, thermal inertia and spin velocity, uniform spin orientation. Simulation includes Venus—Neptune. As orbits evolve, some approach J7:2/M5:9 resonance; there, some get into planet-crossing orbits, hit the Sun, or leave the solar system. Repeat for 750 10-km test asteroids; find fraction that are trapped long enough to reach Mars-crossing orbit.

- Estimate rate of planetary impacts (using different dynamical approximations) from 9024 test bodies in the J7:2/M5:9 resonance.

- Estimate the initial size frequency distribution of the BAF using smooth particle hydrodynamic models of collisions with assumed diameters, compositions, velocities, and collision angles.

not done yet...

- Use estimated initial distribution and a depletion rule to estimate the number of objects of each size in the BAF as a function of time since the collision that formed the family.
- Use the Monte Carlo simulations with test objects to invent a probability distribution for BAF asteroids of a given size to reach the J7:2/M5:9 resonance at time t.
- Use the Monte Carlo simulations with test objects to invent a probability distribution for objects in the resonance to hit Earth at time t after reaching the resonance.
- Finding: 40% chance of no impact from a K/T-size object. 60% chance of one or more impacts in the 160 My since the inferred origin of the BAF.

60%? I thought they said >90%?
(and what about 160My vs. 65My?)

There's still more: Bayes calculation depending on background rates... NEO and comet impact rates...

So, where does the randomness come from?

Ultimately, it comes from the Gaussian, uniform, and truncated Gaussian distributions used in the Monte Carlo simulations, plus independence assumptions, etc.

Does not include observational errors, catalog issues, or similar (as far as I can tell).

Even if you buy all the approximations to the dynamics, uncertainties in the catalogs, etc., the stochastics seem tenuous.

Where does probability come from in IP?

- The state of Nature can be thought of as random (subjective [Bayesian] approach)

  According to I.J. Good (1965),

  "...the essential defining property of a Bayesian is that he regards it as meaningful to talk about the probability P(H|E) of a hypothesis H, given evidence E."

- Observational errors can be modeled as random (both frequentists and Bayesians)
- The physical process can be modeled as random (Big-Bang cosmology, earthquakes, ...) Models need to be calibrated and tested.

Technical Stuff—Outline

- Inverse Problems as Statistics
  - Ingredients; Models
  - Forward and Inverse Problems—applied perspective
  - Statistical point of view
  - Some connections
- Linear inverse problems
- Qualitative uncertainty: Identifiability and uniqueness
  - Sketch of identifiablity
  - Example: interpolation with systematic error & noise

Technical outline, contd.

- Quantitative uncertainty: Decision Theory
  - Comparing decision rules: Loss and Risk
  - Strategies. Bayes/Minimax duality
  - Example: Shrinkage estimators and MSE Risk
  - Illustration: Estimating a bounded normal mean

Inverse Problems as Statistics

- Measurable space X of possible *data* X.
- Set Θ of descriptions of the world—*models*.

Typically Θ has special structure.

- Family $\mathsf{P} = \{\mathsf{P}_\theta : \theta \in \Theta\}$ of probability distributions on X indexed by models θ.
- *Forward operator* $\theta \to \mathsf{P}_\theta$ maps model θ into a probability measure on X .

X –valued data X are a sample from P_θ.

P_θ is all: randomness in the "truth," measurement error, systematic error, censoring, *etc*.

Inverse Problems

Observe data X drawn from P_θ for some unknown $\theta \in \Theta$. (Assume Θ contains at least two points; otherwise, data superfluous.)

Use X and the knowledge that $\theta \in \Theta$ to learn about θ; for example, to estimate a *parameter* $g(\theta)$ (the value $g(\theta)$ at θ of a G-valued function g defined on Θ).

Forward Problems in Science & Engineering

Often thought of as a composition of steps:

- transform idealized model θ into perfect, noise-free, infinite-dimensional data ("approximate physics")
- keep a finite number of the perfect data, because can only measure, record, and compute with finite lists
- possibly corrupt the list with measurement error.

Equivalent to single-step procedure that includes censoring and corruption.

Inverse Problems in Science & Engineering

Inverse problems often “solved” using applied math methods for Ill-posed problems (*e.g.*, Tichonov regularization, analytic integral or differential inversions)

Those methods are designed to answer different questions; can behave poorly with data (*e.g.*, bad bias & variance)

Inference ≠ construction: Statistical viewpoint may be better for real data with random errors.

Elements of the Statistical View

Distinguish between characteristics of the problem and characteristics of the methods.

Identifiability is a fundamental qualitative property of a parameter:

g is *identifiable* if for all $\eta, \zeta \in \Theta$,

$$\{g(\eta) \neq g(\zeta)\} \Rightarrow \{P_\eta \neq P_\zeta\}.$$

In most inverse problems, $g(\theta) = \theta$ isn't identifiable, nor are most linear functionals of θ.

Deterministic/Statistical Connections

Identifiability—distinct parameter values yield distinct probability distributions for the observables— is similar to *uniqueness*—forward operator maps at most one model into the observed data.

Consistency—parameter can be estimated with arbitrary accuracy as the number of data grows—is related to *stability* of a recovery algorithm—small changes in the data produce small changes in the recovered model.

∃ quantitative connections too.

Linear Forward Problems

A forward problem is *linear* if

- Θ is a subset of a separable Banach space T
- $X = \Re^n$, $X = (X_j)_{j=1}^{n}$
- For some fixed sequence $(\kappa_j)_{j=1}^{n}$ of elements of T* (the normed dual of T),

$$X_j = \langle \kappa_j, \theta \rangle + \varepsilon_j, \quad \theta \in \Theta,$$

where $\varepsilon = (\varepsilon_j)_{j=1}^{n}$ is a vector of stochastic errors whose distribution does not depend on θ.

Linear Forward Problems, contd.

Linear functionals $\{\kappa_j\}$ are the "representers"

Distribution P_θ is the probability distribution of X.

Typically, $\dim(\Theta) = \infty$; at least, $n < \dim(\Theta)$, so estimating θ is an underdetermined problem.

Define

$$K : T \rightarrow \Re^n$$

$$\theta \mapsto (<\kappa_j, \theta>)_{j=1}^{n} .$$

Abbreviate forward problem by $X = K\theta + \varepsilon$, $\theta \in \Theta$.

Linear Inverse Problems

Use $X = K\theta + \varepsilon$, and the constraint $\theta \in \Theta$ to estimate or draw inferences about $g(\theta)$.

P_θ, the probability distribution of X, depends on θ only through $K\theta$, so if there are two points

$$\theta_1, \theta_2 \in \Theta \text{ such that } K\theta_1 = K\theta_2 \text{ but}$$

$$g(\theta_1) \neq g(\theta_2),$$

then $g(\theta)$ is not identifiable.

Example: Interpolation w/ systematic error & noise

Observe $X_j = f(t_j) + \rho_j + \varepsilon_j, \quad j = 1, 2, \ldots, n.$

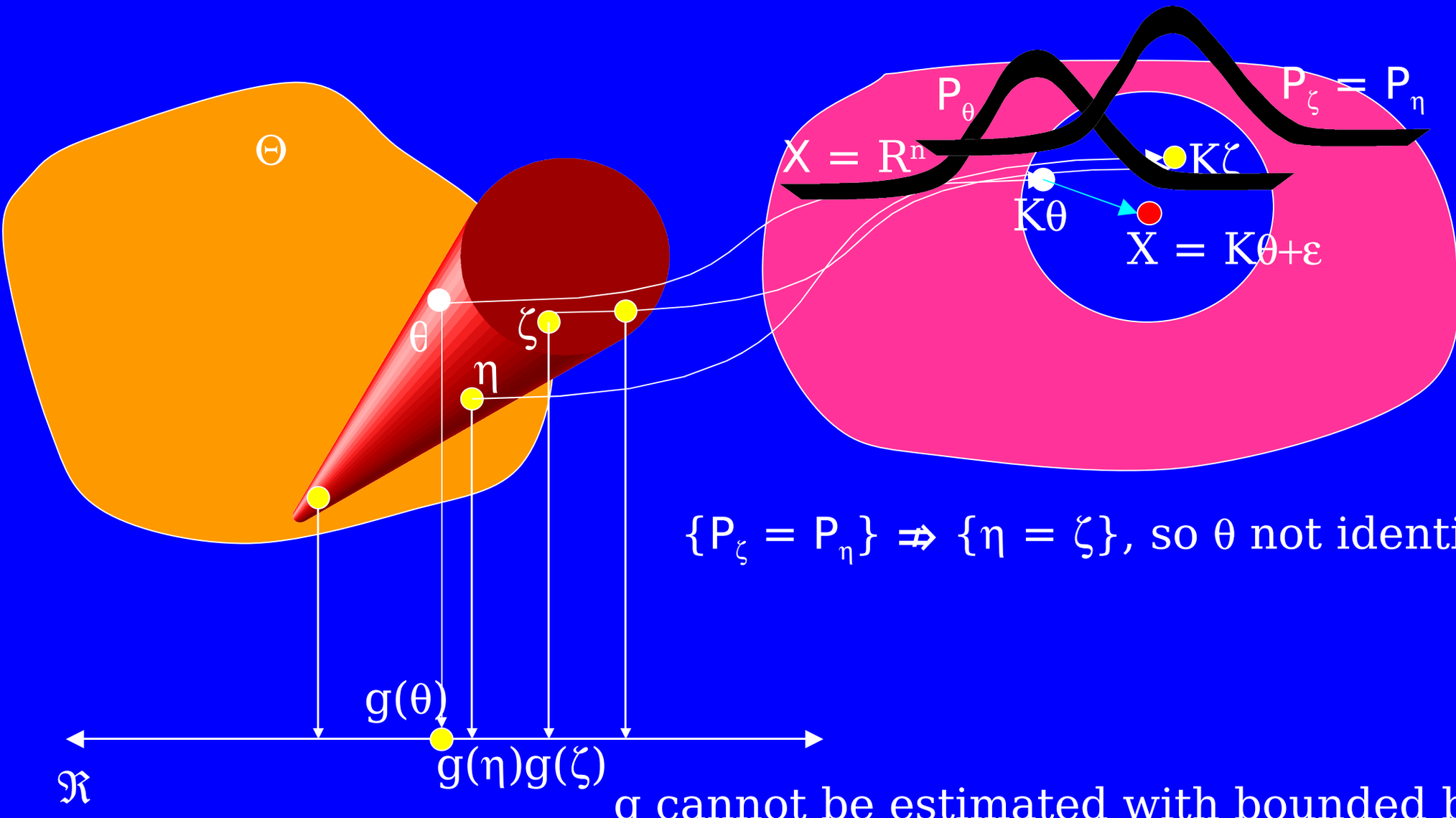
- $f \in C$, a set of smooth of functions on $[0, 1]$
- $t_j \in [0, 1]$
- $|\rho_j| \leq 1$, $j=1, 2, \ldots, n$
- ε_j *iid* $N(0, 1)$

$\Theta = C \times [-1, 1]^n$, $X = \Re^n$, and $\theta = (f, \rho_1, \ldots, \rho_n)$.

Then P_θ has density

$$(2\pi)^{-n/2} \exp\{-\Sigma_{j=1}^{n} (x_j - f(t_j) - \rho_j)^2\}.$$

Sketch: Identifiability



$\{P_\zeta = P_\eta\} \nRightarrow \{\eta = \zeta\}$, so θ not identifiable

g cannot be estimated with bounded bias

$\{P_\zeta = P_\eta\} \nRightarrow \{g(\eta) = g(\zeta)\}$, so g not identifiable

Loss and Risk

- 2-player game: Nature v. Statistician.
- Nature picks θ from Θ.
  θ is secret, but statistician knows Θ.
- Statistician picks δ from a set D of rules.
  δ is secret.
- Generate data X from P_θ, apply δ.
- Statistician pays *loss* $L(\theta, \delta(X))$. L should be dictated by scientific context, but...
- *Risk* is expected loss: $r(\theta, \delta) = E_\theta L(\theta, \delta(X))$
- Good rule δ has small risk, but what does *small* mean?

Strategies

Rare that one δ has smallest risk $\forall\theta\in\Theta$.

- δ is *admissible* if no estimator does at least as well for every θ, and better for some θ.
- *Minimax decision* minimizes

  $r_\Theta(\delta) \equiv \sup_{\theta\in\Theta} r(\theta, \delta)$ over $\delta\in D$ (Nature picks θ cleverly)
- *Minimax risk* is $r_\Theta^* \equiv \inf_{\delta\in D} r_\Theta(\delta)$
- *Bayes risk* of δ for *prior probability distribution* π is

  $r_\pi(\delta) \equiv \int_\Theta r(\theta,\delta)\pi(d\theta)$ (Nature picks θ at random from π)
- *Bayes decision* minimizes $r_\pi(\delta)$ over $\delta\in D$
- *Bayes risk* is $r_\pi^* \equiv \inf_{\delta\in D} r_\pi(\delta)$.

Minimax often Bayes for *least favorable prior*

Generally for convex $\Theta, D,$ concave-convexlike $r,$

$$\inf_{\delta\in\mathcal{D}}\sup_{\theta\in\Theta} r(\theta,\delta) = \sup_{\pi\in\Pi}\inf_{\delta\in\mathcal{D}}\int_{\Theta} r(\theta,\delta)d\pi(\theta).$$

I.e.,

$$r^*_{\Theta} = \sup_{\pi\in\Pi} r^*_{\pi}.$$

If minimax risk $\gg$ Bayes risk, prior π (not data and constraints) controls the apparent uncertainty of the Bayes estimate.

Example: Bounded Normal Mean

Observe $X \sim N(\theta, 1)$. Know *a priori* $\theta \in [-\tau, \tau]$.

Want to estimate $g(\theta) \equiv \theta$.

Squared-error loss:

$$L(\theta, \delta) = (\theta - \delta)^2$$

$$r(\theta, \delta) = E_\theta L(\theta, \delta(X)) = E_\theta (\theta - \delta(X))^2$$

$$r_\Theta(\delta) = \sup_{\theta \in \Theta} r(\theta, \delta) = \sup_{\theta \in \Theta} E_\theta (\theta - \delta(X))^2$$

$$r_\Theta^* = \inf_{\delta \in D} \sup_{\theta \in \Theta} E_\theta (\theta - \delta(X))^2$$

Risk of X for bounded normal mean

Naive (& maximum likelihood) estimator is

$$\delta(X) \equiv X.$$

$EX = \theta$, $\therefore$ X unbiased for θ, $\therefore$ θ unbiasedly estimable.

$$r(\theta, X) = E_\theta\,(\theta - X)^2 = \mathrm{Var}(X) = 1.$$

Consider uniform prior to capture constraint $\theta \in [-\tau, \tau]$:

$\theta \sim \pi = U[-\tau, \tau]$ = uniform distribution on $[-\tau, \tau]$.

$$r_\pi(X) = \int_{-\tau}^{\tau} r(\theta, X)\,\pi(d\theta) = \int_{-\tau}^{\tau} (2\tau)^{-1}\,d\theta = 1.$$

Frequentist risk of X equals Bayes risk of X for uniform prior π (but X is not the Bayes estimator).

Truncation is better (but not best)

Easy to find an estimator better than X from both frequentist and Bayes perspectives.

Truncation estimate δ_T

$$\delta_T(x) \equiv (x \vee -\tau) \wedge \tau = \begin{cases} -\tau, & x \leq -\tau \\ x, & -\tau < x < \tau \\ \tau, & x \geq \tau. \end{cases}$$

δ_T is biased, but has smaller MSE than X, $\forall \theta \in \Theta$.

Minimax MSE Estimate of BNM

Truncation estimate better than X, but neither minimax nor Bayes.

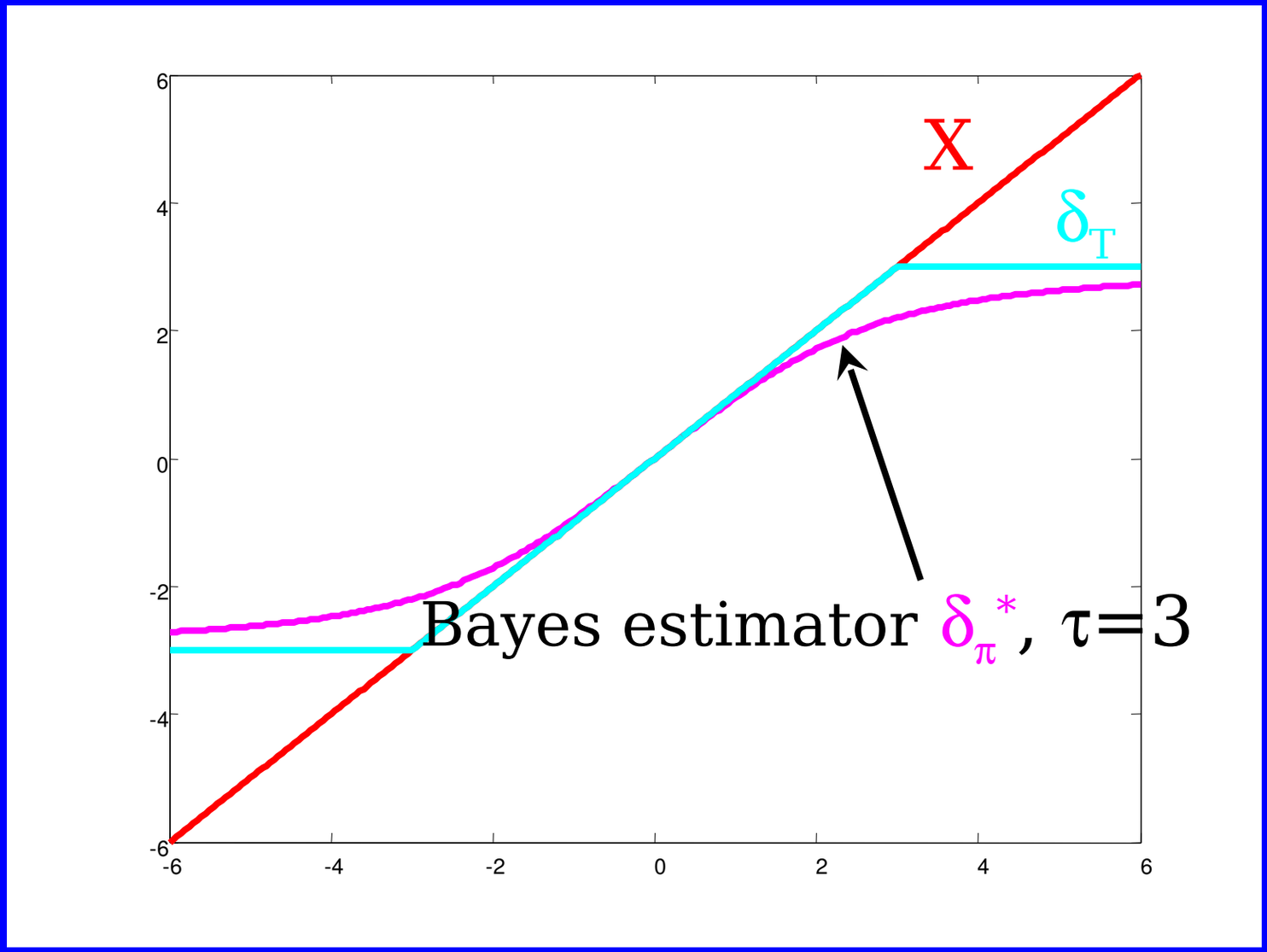
Clear that $r^* \leq \min(1, \tau^2)$ since MSE(X) = 1, and $r_\Theta(0) = \tau^2$.

Minimax MSE estimator is a nonlinear *shrinkage estimator*.

Minimax MSE risk for affine estimators is $\tau^2/(1+\tau^2)$.

Minimax MSE for nonlinear estimators no less than 4/5*(minimax affine risk)

Bayes estimator is also nonlinear shrinkage



δ_π^*

For $\tau = 3$, Bayes risk $r_\pi^* \approx 0.7$ (by simulation) .

Minimax risk $r_\Theta^* = 0.75$.

Bayes & Minimax Risks

τ	r_π^* **(simulation)**	r_θ^* **(lower bound)**
0.5	0.08	0.16
1	0.25	0.40
2	0.55	0.64
3	0.70	0.72
4	0.77	0.75
5	0.82	0.77
>>1	→1	→1

Difference between knowing $\theta \in [-\tau, \tau]$ and $\theta \sim U[-\tau, \tau]$.
(3^{rd} column is a lower bound, 4/5*(minimax affine risk)

Bayes & Minimax Risks

Difference between knowing $\theta \in [-\tau, \tau]$ and $\theta \sim U[-\tau, \tau]$.
(last column is a lower bound, 4/5*[minimax affine risk])

Summary

- To quantify uncertainty statistically/probabilistically, the randomness has to come from somewhere. Where?
  State of nature (Bayesian)?
  Noise (Bayesian and frequentist)?
- Statistical viewpoint is useful abstraction.
  Physics in mapping $\theta \mapsto P_\theta$
  Prior information in constraint $\theta \in \Theta$.
- There is more information in the assertion $\theta \sim \pi$, with π supported on Θ, than there is in the constraint $\theta \in \Theta$.
- Separating "model" θ from parameters $g(\theta)$ of interest is useful. Many interesting questions can be answered without estimating the entire model.
- Thinking about measures of performance is illuminating.

References

Bottke, W.F., D. Vokrouhlický & D. Nesvorný, 2007. An asteroid breakup 160 Myr ago as the probable source of the K/T impactor, *Nature*, **449**, 48–53.

Evans, S.N. & Stark, P.B., 2002. Inverse Problems as Statistics, *Inverse Problems*, *18*, R1-R43.

Evans, S.N., B. Hansen, and P.B. Stark, 2005. Minimax Expected Measure Confidence Sets for Restricted Location Parameters, Bernoulli, 11, 571-590.

Freedman, D.A. and P.B. Stark, 2003. What is the Chance of an Earthquake? in *Earthquake Science and Seismic Risk Reduction*, F. Mulargia and R.J. Geller, eds., NATO Science Series IV: Earth and Environmental Sciences, v. 32, Kluwer, Dordrecht, The Netherlands, 201-213.

Littlewood, J.E., 1953. *A Mathematician's Miscellany*, Methuen & Co., Ltd., London, 136pp.

Schafer, C.M. and P.B. Stark, 2004. Using what we know: inference with physical constraints. Proceedings of the Conference on Statistical Problems in Particle Physics, Astrophysics and Cosmology PHYSTAT2003, L. Lyons, R. Mount and R. Reitmeyer, eds., Stanford Linear Accelerator Center, Menlo Park, CA, 25-34.

Stark, P.B., 1992. Inference in infinite-dimensional inverse problems: Discretization and duality, *J. Geophys. Res.*, *97*, 14,055-14,082.